Spatial Demography ***2013*** 1(1): 41-55
http://spatialdemography.org

SPATIAL DEMOGRAPHY
ISSN 2164-7070 (online)

RESEARCH

Public Use Microdata Area Fragmentation: Research and Policy Implications of Polygon Discontiguity

Carlos Siordia[a] and Amber Fox[b]
[a]University of Texas Medical Branch
[b]Texas A&M University

ABSTRACT

The American Community Survey (ACS) is created by the Unites States Census Bureau and contains the most recent and detailed information on the population. The U.S. Census Bureau releases Public Use Microdata Sample (PUMS) ACS files to allow public entities the ability to examine detailed individual-level data. In order to protect the confidentiality of survey participants, the Census only allows individual-level observations to be geographically referenced with Public Use Microdata Areas (PUMAs)—polygons that contain at least 100,000 people. Microdata files offer the unique advantage of creating custom tabulations on the population that are not made available elsewhere. The advantages from using PUMS files are limited by the typology of PUMAs. Our project fills a gap by introducing the reader to how PUMA typology is produced and how it affects research. The specific aim of this project is to highlight the fact that PUMAs can be fragmented—made up of physically non-adjacent/multi-part polygons. This is important because the treatment of fragmented/multi-part polygons as contiguous spatial entities erroneously imposes a false structure of contiguity that challenges theoretical and statistical assumptions in geographically aware research. By providing evidence from three metropolitan areas in Texas (USA) and displaying the degree of polygon discontiguity concentration by state, we outline some of the implications of polygon discontiguity for research and policy. We compliment our descriptive investigation by discussing solutions in closing and by identifying and discussing the three general elements in polygon fragmentation: quantity, distance, and size.

KEYWORDS: GIS, PUMS, PUMA, ACS, microdata, spatial demography

INTRODUCTION

The proliferating production and consumption of spatially referenced data has the potential of being misused (Grubesic and Matisziw 2006). For example, one of the limitations when working with spatial data is geospatial mismatch (Jacques 2004). Unlike discussions of the "spatial mismatch" arising from living in one location while having a set of work skills that requires employment be sought at a different location (see Arnott 2008; Kain 1968), "geospatial mismatch" occurs when geographically referenced data does not match its assigned location in spatial analysis. For example, taking information from a sub-county

[a]University of Texas Medical Branch

Corresponding Author: Carlos Siordia, University of Texas Medical Brach, Sociomedical Division, 301 University Blvd, Galveston, TX, 77555
E-mail: csiordia@gmail.com

polygon and assigning it to the whole county. The generalization of information from the sub-county to the full-county polygon imposes a false location for the measured spatial attribute—an event that may have serious implications in the interpretation of analytical results.

Although technical in nature, a major issue in spatial demography is that of computational geometry—the study of algorithms measuring geometric objects (de Berg et al 2008). For example, quantitative approaches measuring spatial nonstationarity (the fact that statistical relationships can vary as a function of geographical location) usually make use of computational algorithms to determine a polygon's geometric centroid to create an adjacency matrix used in the representation of spatial relationships between polygons (Mitchell 2005). Some researchers may not be aware of the fact that geometric centroids are calculable for both *single-part* and *multi-part* polygons.

In contrast to a single-part polygon, where a contiguous geometric-shape is stored as a single feature in a vector file, a multi-part geometric figure contains multiple and geographically separated polygons that are stored as a single feature in a vector file. Since spatial statistical techniques assume no geospatial mismatch (i.e., they rely on the accurate representation of polygons), the presence of geospatial mismatches has the potential for compromising results. The specific aims of this article are to discuss and present evidence of multi-part polygons. Implications for empirical research and theoretical assumptions are discussed throughout. We hope this is a first step towards the development of a full treatise where the effects of multi-part polygons on empirical analysis are delineated.

The use of U.S. Census Bureau microdata is frequented by investigators and policy makers to assess, manage, and allocate governmental funds and services (see Reamer 2010). The Census only maintains complete and unadulterated microdata files for their internal use. The Census does make a version of their internal microdata files available to the public and releases them through Public Use Microdata Sample (PUMS) data files. In order for the Census to make a version of the internal microdata file public, they must undertake several procedures to ensure the confidentiality of survey respondents when releasing the PUMS files. One of the procedures is to only allow public data users the ability to physically locate a "micro-level unit" (i.e., a person-unit record) to a geographical area (a macro-level unit) with at least 100,000 people. These geographical polygons are referred to as Public Use Microdata Areas (PUMAs).

There are many approaches employed in investigations which seek to make use of geographically referenced data. For example, hierarchical models (Raudenbush et al. 2002) and spatial approaches (Fotheringham et al. 2002) have been employed to make use of geographically referenced data. Such techniques are engaged so as to explore how macro-level factors play a role in micro-level outcomes. There are four popular instances where researchers are in need of using PUMAs: (1) where the micro-unit is the unit of analysis in a model that includes macro-unit measures; (2) where the micro-unit is the unit of analysis and its geographical location is required in the estimation of the model; (3) where the macro-unit, as the unit of analysis, requires a customized estimate; and (4) where the PUMA polygon is used in areal interpolation procedures.

In the first case, investigators may at times be in need of nesting individuals into geographies with known characteristics. By nesting we simply mean that micro-level units (i.e., people) are linked with macro-level units (i.e., PUMAs). For example, in recent years, social scientists investigated the "language shift" of second- and third-generation children of immigrants (Alba et al. 2002). The authors decided to include a social-environmental measure because they thought the level of "concentration of mother-tongue speakers" in the person's area of residence could play a role in whether the children of immigrants retained their mother-tongue. Because they used 1990 PUMS data, they were forced to account for their "proximity

to group residential concentration" using PUMA-level measures. That is, individual-level units were assigned a level of concentration of mother-tongue speakers according to their PUMA of residence. After following these procedures and executing their models, the authors explained their finding that for some children of immigrant groups, "communal context" plays an important role in language shift—where those living in highly concentrated areas of co-mother-tongue speakers experience less of a language shift. In closing, they do make it clear that the "geographic effect" deserves more analysis than is given in their article (Alba et al. 2002:480). A policy investigation also made use of PUMAs to define central city-suburban boundaries, and found that "within suburbs, the poor generally live in communities that have somewhat below-average number of jobs" (Raphael and Stoll 2010).

In the second instance, the micro-unit of analysis and its geographical location is required in the estimation of a model. For example, in a recent and mathematically sophisticated publication, economists explored how the structure of the error covariance matrix correlated in magnitude within and between spatial clusters at the state, division, and PUMA levels (Barrios et al. 2010). They explore an individual's characteristic (like years of education) while accounting for their geographical location in relation to other people—where latitude and longitude measures are assigned using PUMA polygons. In their explanations, they show how a PUMA's geographical location plays a role in their estimation of spatial clustering. They eventually provide evidence for how ignoring spatial correlation (at a sub-state level such as PUMAs) in outcomes (such as years of education) may bias standard errors. In particular, the authors report findings that within-PUMA correlations are larger than within-state correlations—meaning that PUMA-level spatial autocorrelation investigations merit attention. This is why they conclude by recommending that others should assess the extent to which spatial correlation at a sub-state level is playing a role in the micro-level outcome of interest (Barrios et al. 2010). There are also policy investigators, such as the Working Poor Families Project (WPFP), who use PUMS data to display customized estimates by PUMAs (Rivers 2006).

In the third instance, the macro-unit is the unit of analysis and its measure requires customized tabulations that can only be attained from using microdata. Of particular importance here is the need for a "customized" estimate. For example, if a researcher wants to conduct a spatial clustering analysis on county-level estimates on people under age 18, the Census provides those estimates by county in public tabulations.[1] There are "already made" and public estimates which can be geographically referenced. However, some macro-level measures may not be readily available in Census provided tabulations that can be geographically referenced (i.e., geographically located). Researchers would then customize macro-level measures—requiring the use of microdata unit characteristics and their sample weights to compute estimates. In such a case, public data users would be relegated to using PUMS data with PUMA geographies.

For example, a geographer used PUMAs to investigate the spatial patterns of ethnic integration in the US (Wong 1998). He created a custom estimate on the percentage of "multiethnic households" at the PUMA-level—he argues this is a good measure of ethnic integration. Wong (1998) then goes on to show that "ethnic integration" varies widely by PUMA geography. In a more recent study, a sociologist used PUMAs as the unit of analysis (in a geographically weighted regression: GWR) to investigate the macro-level association between Latino population concentrations and the number of people in poverty within the PUMA (Siordia and Farias 2013). This research made use of PUMS data in order to account for a customized "percent of non-Latino-Black" GWR-coefficient by PUMA polygon. Since it is preferable to use the smallest level of geography available in order to obtain segregation measures and contextual characteristics that accurately capture the experiences of individuals in their residential locations, researchers who study residential attainment and segregation have also used PUMAs as the level of analysis when they are limited to using PUMS data. For

example, Yu and Myers (2007) utilized PUMAs in their analysis of immigrant "residential assimilation," developing measures of percent white, percent co-ethnics, and median income at the PUMA-level. They then investigated how individual social characteristics were related to the composition of their "neighborhood" (i.e., PUMA).

In the fourth and final instance, PUMA polygons are necessary for areal interpolation research. In a recent study contrasting four data reassignment procedures, the use of areal interpolation through population weighting (with block-level census data) is presented (Saporito et al. 2007). In their elegant study, the authors interpolate with areal weighting, "assigning per-pupil expenditures from school districts to 2000" PUMAs. This was done "to explore how school district funding was related to private school attendance" (Saporito et al. 2007:888). In particular, they "interpolate the percent of white students living in all US school districts to all" PUMA areas (Saporito et al. 2007:889) and found that when they "produced an estimate of the percent of white children in each PUMA" the population weighting method was the most efficient (Saporito et al. 2007:915). Policy researchers developing the Educational Needs Index (ENI) 2.0 have made use of PUMAs. The ENI is a study on the pressures that influence educational policy and planning.[2]

As is made clear from our preceding discussion, researchers and policy makers are at times in need of investigating micro-level or macro-level phenomenon that requires geographically referenced data where customized context-measures are necessary. Many of them do not have access to the internal and highly guarded microdata that could potentially allow them to physically locate individuals down to their place of residence. This presents a challenge with a spatial dimension. When the topic of interest requires that investigators use geographically referenced microdata, they must either create their own geographically referenced data or use secondary data sources. Since many lack the funds required to conduct large scale studies or are involved in time-sensitive projects, the use of secondary data sources (like PUMS files) is an attractive alternative.

The ever expanding research field that employs geographically referenced data faces many theoretical and methodological challenges. Of particular importance here are the theoretical premises underlying the geoboundarization (i.e., the act of delineating the geographic boundaries of a place) of macro-level units. This research is propelled by the idea that a person's context matters (Siordia 2011). In other words, the motivation for including customized macro-level measures in analysis is in part based on the assumption that everything is related to everything else, but near things are more related than distant things (Tobler 1970). Under such a theoretical position, the contiguity of the geographical polygon is an implicit *and* necessary condition. By contiguity, we refer to how geographical units can have a spatial "property of sharing a common boundary or vertex" (Grubesic and Matisziw 2006) and by noncontiguity we mean that polygons can be made up of multiple parts that neither shares a boundary nor vertex. For example, if a person wanted to travel to all the internal points in a *noncontiguous* polygon (i.e., a multi-part polygon), he/she would have to exit the polygon at some point to enter it again in a different location (Cova and Church 2000). Thus, when we say a polygon is *fragmented*, we mean that it is made up of multiple parts that are not physically connected at any one point or line.

To make it clear, our core argument is that *polygon contiguity is necessary* in investigations which seek to either: (1) account for macro-level measures because they are perceived as having a significant role in micro-level outcomes; and (2) explore how statistical relationships vary as a function of geographical location. Most social science research assumes the fundamental polygon typology of contiguity, so that the "lack of spatial contiguity can have a dramatic impact on spatial statistical analysis" and theory (Grubesic and Matisziw 2006). The treatment of fragmented polygons as single geographical units imposes a false structure of contiguity that may create systemic theoretical and

statistical errors, because although subtle and primarily implicit, these geo-spatially aware approaches are founded on the theoretical axiom that both micro- and macro-level processes are influenced by geographical location—and embedded deeply in this assertion is the fundamental premise that the measurement of location occurs using decipherable *and* contiguous geographical polygons.

There is, to our knowledge, no existing publication where public use microdata from the US Census (and its PUMAs) are used that addresses a peculiar question: How does polygon discontiguity play a role in the investigation of geographically referenced data? This question may not be in the minds of most Census microdata users because many ignore the fact that *polygon discontiguity is present in some Census geographies*. More to our point, many PUMS data users are completely unaware that some PUMAs are geographically fragmented (i.e., are made up of noncontiguous/multi-part polygons). We fill the literature gap by addressing the geographical discontiguity problem present in PUMAs, and in doing so, we outline the three basic components to consider when evaluating geographical fragmentations. The three basic elements of polygon discontiguity discussed are: (1) the *size* of each fragmentation; (2) the *number* of fragmentations; and (3) the *distance* between fragments. We compliment this discussion by diagnosing state-level PUMA fragmentation concentrations with an index we created.

This is the first paper ever published addressing PUMA fragmentation. We illustrate our basic and descriptive findings by making use of PUMAs in three Texas counties: Bexar County, Harris County, and Tarrant County. Figures displaying PUMA fragmentations within these counties are provided to help meet our specific aim of visually displaying for the readers three cases where polygon discontiguity is present. Along with the figures, we introduce the three basic elements of polygon discontiguity. In closing, we discuss possible solutions to this research problem. This descriptive paper is important on theoretical and methodological grounds as spatiologists (i.e., individuals who investigate humans through a geographically aware prism) in the public sector continue exploring human behavior and offering solutions for the social challenges we face.

We describe our study areas primarily on demographic grounds. Bexar County (km^2=3,256) has a population of about 1.7 million people—making it the 19th most populated county in the nation. Bexar County was created in 1836, when Texas reached its statehood. Latinos make up about 55% of its population, and it has a 531/km^2 population density. Bexar County makes up most of the San Antonio metropolitan area. Harris County (km^2=4,605) contains about 4.1 million people, making it the 3rd most populous county in the US and has a 914/km^2 population density. The county was initially founded as Harrisburg County and officially changed its name three years later in 1839. Latinos make up 41% of the population and the county makes up a large part of the Houston metropolitan area. Tarrant County (km^2=2,324) has a population of 1.8 million, which makes it the 16th most populous county in the US. It makes up a part of the Dallas-Fort Worth metropolitan area, Latinos make up 20% of its population, and it has an 809/km^2 population density.

MATERIALS AND METHODS

Although we do provide a state-level rate of PUMA-fragmentation, this study does not consist of outlining all the sub-state areas where PUMA fragmentations occur. We abstain from conducting quantitative analyses of PUMA polygons. Instead, in this descriptive investigation, we introduce the reader to polygon discontiguity in PUMAs by making use of three counties that provide evidence of PUMA fragmentation. In closing, we provide suggestions for future research which would undertake a full investigation on how polygon discontiguity affects quantitative research employing PUMA referenced PUMS data. In this section, we provide sufficient details to allow the replication of this descriptive work.

American Community Survey (ACS) Public Use Microdata Sample (PUMS) files contain

individual-level records on the characteristics of the US population.[3] ACS data is important because it influences the allocation of almost 70% of all federal grant funding and is directly responsible for the yearly distribution of more than 80 billion federal dollars (Reamer 2010). Additionally, all local governments (e.g., state, school districts) *must* use ACS data to challenge any population estimates they disagree upon with the federal government.[4] For example, Small Area Income and Poverty Estimates (SAIPE) are calculated using ACS data and influence the funding schools receive. If a local government disagrees with the SAIPE estimates they can only challenge the federal government by making use of ACS microdata.[5]

PUMS files were first created after the 1960 Decennial Census and continue to date.[6] As mentioned before, PUMAs are unique statistical geographical areas used by the US Census Bureau in the public dissemination of internal microdata (i.e., PUMS files) that ensure the confidentiality of survey participants. PUMAs only respect state boundaries, are non-overlapping, and contain a decennial Census population of at least 100,000 or more. Beginning in 1990, State Data Centers (SDCs) started collaborating with federal, regional, state, and local agencies to help delineate the geographical boundaries of PUMA polygons.[7]

The US Census Bureau outlines basic criteria and suggestions as they work with their "local" SDC partners—who play a key role in the final decision of their states' PUMA "boundarizations". The PUMAs in our study use 2000 boundaries because the ones from the 2010 decennial Census will not be available until sometime in 2012. The history of the relationship between the Census and SDCs—as it relates to PUMA "geo-boundarization"—is long and complex.[8] The short version of the story is that in 1990 the process for PUMA delineation was very flexible and vague. By 2000, it had improved somewhat, but the process was made much clearer during the 2010 decennial census. For both 1990 and 2000, the Census only provided "guidelines" for how SDCs should delineate their PUMA boundaries. By guidelines, we mean that SDCs were not obligated to follow the "suggestions" offered by the Census on how PUMAs should be delineated.

It was not until 2010 that the Census, for the first time, introduced "criteria" for delineating the polygons. By criteria we mean that SDCs were obligated to follow a particular list of procedures. For PUMAs to be released in 2012, SDCs: (1) were only given the opportunity to *suggest* PUMA boundaries; (2) were *obligated* to follow Census Bureau "criteria" in delineating PUMAs; (3) even if they deviated from the "guidelines," they were required to provide an explanation and even then the Census would only take their non-guideline compliance polygons under consideration; and (4) regardless of the steps taken in any of these circumstances, the Census would check every SDC-delineated PUMA boundary to ensure they followed both the criteria *and* guidelines set out from them. Because of these procedures, 2010 PUMA boundaries, compared to our 2000 polygons, may contain less fragmentation.

In all the decades, SDCs were seen as the Census' participants in the PUMA boundary delineation process. However, beyond this relationship and throughout the decades, procedures used by states to geo-boundarize PUMAs (i.e., delineate the geographical boundaries) have differed—with some of the more populous states farming out the work to county governments, regional and/or metropolitan planning agencies, other state agencies, and/or large city planning departments[5]. After the SDCs' partners had developed the boundaries, the SDCs would then organize their work and submit them to the Geography Division at Census Bureau HQ in Suitland, Maryland. The 2000 PUMA polygons in our study are the product of ambiguous processes born out of the flexible relationship between the Census, their SDC partners, and their sub-SDC participants.

The Census criterion for PUMA delineation has changed over the years. In 2000[9], counties, minor civil divisions, incorporated places, and

tracts were used in the geoboundarization of PUMA[10] polygons. By 2010, counties and census tracts were the only PUMA building block geographies used.[11] PUMA boundaries are most heavily influenced by population distribution, building block geographical entities, and contiguity criteria. The key element here is what is meant by "contiguity." In publicly available documentation[4], the Bureau reports that each PUMA must constitute a geographically contiguous area. The Bureau defines *contiguous areas* as geographical entities that share common boundaries and considers *noncontiguous areas* to be formed by disjointed pieces. Other Census publications have further explained that contiguity is present if adjoining building block polygons are connected by a(.*n*) line(s) or area(s) and that their boundaries are said to be noncontiguous if they do not touch or are only connected at a corner.[12] In other words, PUMA geographic building blocks are noncontiguous if they do not connect along at least one polygon line.

This basic and logical "contiguity requirement" has an exception which is largely ignored by both academic and non-academic data users. For example, a 2010 PUMA is allowed to be *noncontiguous* if a county or census tract is noncontiguous. In more technical terms, the contiguity of a PUMA polygon is determined by the contiguity in its geographical building blocks (i.e., county and tract polygons). The census did reduce the amount of polygon discontiguity by excluding "incorporated places" as building blocks for PUMAs in 2010. However, PUMA geographical discontiguity is not expected to be completely eradicated for 2010 PUMAs and is present in both 1990 and 2000 polygons.

The US Census Bureau releases PUMA geographies using a combination of numeric or alphanumeric codes—from the *geocode* system. We use Topological Integrated Geographic Encoding Referencing (TIGER) Shapefiles from the Census to conduct all of our mapping and spatial analyses. In brief, a shapefile is a popular geospatial vector data format used in geographic information system (GIS) related software. Shapefiles provide open specification for data interoperability—they spatially describe geometries by using points, polylines, and polygons. Full details on ".shp" files are made available elsewhere (ESRI 1998) and a broader explanation of how geographies are released through US Census Bureau TIGER/Line Shapefiles is also available (US Census Bureau 2007). Our 2007 TIGER/Line Shapefiles contain the geographic boundaries as of January 1, 2007, which includes a Census 2000 vintage geography. All analyses and mapping are conducted using ArcGIS® [software by ESRI. ArcGIS® and ArcMap™ are the intellectual property of ESRI and are used herein under license (Copyright © ESRI, all rights reserved) for more information about ESRI® software, please visit www.esri.com] (ESRI 2011).

To diagnose the state-level rate of PUMA-fragmentation, we were inspired by the work of others (Grubesic and Matisziw 2006) and developed the Coefficient of Polygon Fragmentation (CPF_i). CPF_i measures the state-level concentration of polygon fragments. As a diagnostic, the resulting CPF_i values are a representation of polygon fragmentation concentration and thus signal the potential for biased results in empirical analysis that could be associated with the state. The CPF_i is calculated as follows:

$$CPF_i = \frac{x_i / \sum_i^n x_i}{y_i / \sum_i^n y_i;}$$

where x_i is the number of multi-part PUMA polygons in state *i*; and y_i is the number of PUMA polygons in state *i*. The interpretation of CPF_i is as follows: when CPF_i is less than 1, it signals a below average level of PUMA fragmentation in the state; when CPF_i equals 1, it signals an average level of PUMA fragmentation in the state; and when CPF_i is greater than 1, it signals an above average level of PUMA fragmentation in the state.

RESULTS AND DISCUSSION

The primary method for investigating PUMA fragmentation is the presentation of figures below accompanied with a brief discussion highlighting where polygon discontiguity occurs. We start with PUMAs in Bexar County. As shown in Fig 1, we see can that PUMA 5609 (red on the fig) is made up of five non-adjacent polygons. Also, from Figure 1, we are able to detect that PUMA 5610 (blue on the figure) is made up of fourteen non-adjacent polygons. From the two PUMAs, number 5610 is much more fragmented than 5609. Here we point out the first element of polygon discontiguity: the number of fragmentations. When addressing nonadjacent geographical entities, researchers should first question: How fragmented are they? Answering such a question has both statistical and policy implications. For example and in our case, the degree of PUMA fragmentation could reveal how the different processes of the various governmental agencies played a role in creating the issue—with some state governments creating more splits than other states who may be better funded and/or staffed.

From our investigation of Harris County, we find three fragmented PUMAs. PUMA 4619 (red on the figure) is made up of three non-adjacent polygons, while PUMA 4620 (blue in fig) is made up of six polygons, and PUMA 4622 (green in figure) is made up of eight non-adjacent polygons. With this image, we highlight the second element of polygon discontiguity: the distance between fragmentations. In addition to asking about the "degree," researchers should assess the distance encompassed in the non-adjacency. Here again, we argue this question has policy and statistical implications. For example, when conducting spatial analysis, it is common to use the "center of gravity" to assign the polygon's center. In such cases, we could ask: how does the distance between fragmentations affect the "center of the polygon"? It could be that highly dispersed fragmentations create polygon centers outside any of the fragments.

From Figure 3, we can see that within Tarrant County, PUMA 2508 (red in fig) is made up of seven non-adjacent polygons and that PUMA 2509 (blue in fig) is made of nineteen non-adjacent polygons. With this last descriptive picture, we underscore the third and final element of polygon discontiguity: the *size* of each fragmentation. As you can see, PUMA 2509 is made up by some tiny fragments. The size of the fragments may matter and have serious implications for research and policy. For example, after accounting for amount, distance, and size, we could ask a policy related question: How do extremely severed PUMAs reveal local governmental inefficiencies? In the statistical realm, we could ask: How does the inequality in the size of the fragments affect the center of the polygon?

Our descriptive evaluation of three counties clearly shows evidence that PUMA polygon discontiguity is present—this revelation is our substantive contribution. From our previous discussion, it is evident that revealing PUMA fragmentation is important because it has many theoretical and methodological implications for policy and academic researchers. Our work is significant because it introduces, for the first time in a scientific publication, the fact that geographical polygon discontiguity is present in PUMAs. The paper fills a previously unknown gap by expounding how basic theoretical and methodological assumptions are challenged when geographical fragmentation is present.

Table 1 displays CPF_i results. Please note that Oklahoma, Nevada, North Carolina, Texas, and Vermont are all well above the average level of fragmentation. States like Utah and North Dakota have no multi-part PUMA polygons. Texas and California, unlike New York and Florida, are large states with above average multi-part polygon concentrations. There appears to be no clear pattern, detectable through quantitative techniques, of why a state's population size, topology, or cartography would influence the presence of multi-part PUMA polygons. It could be that state-level administrative procedures play

Figure 1. PUMA Fragmentation Results for Bexar County, TX

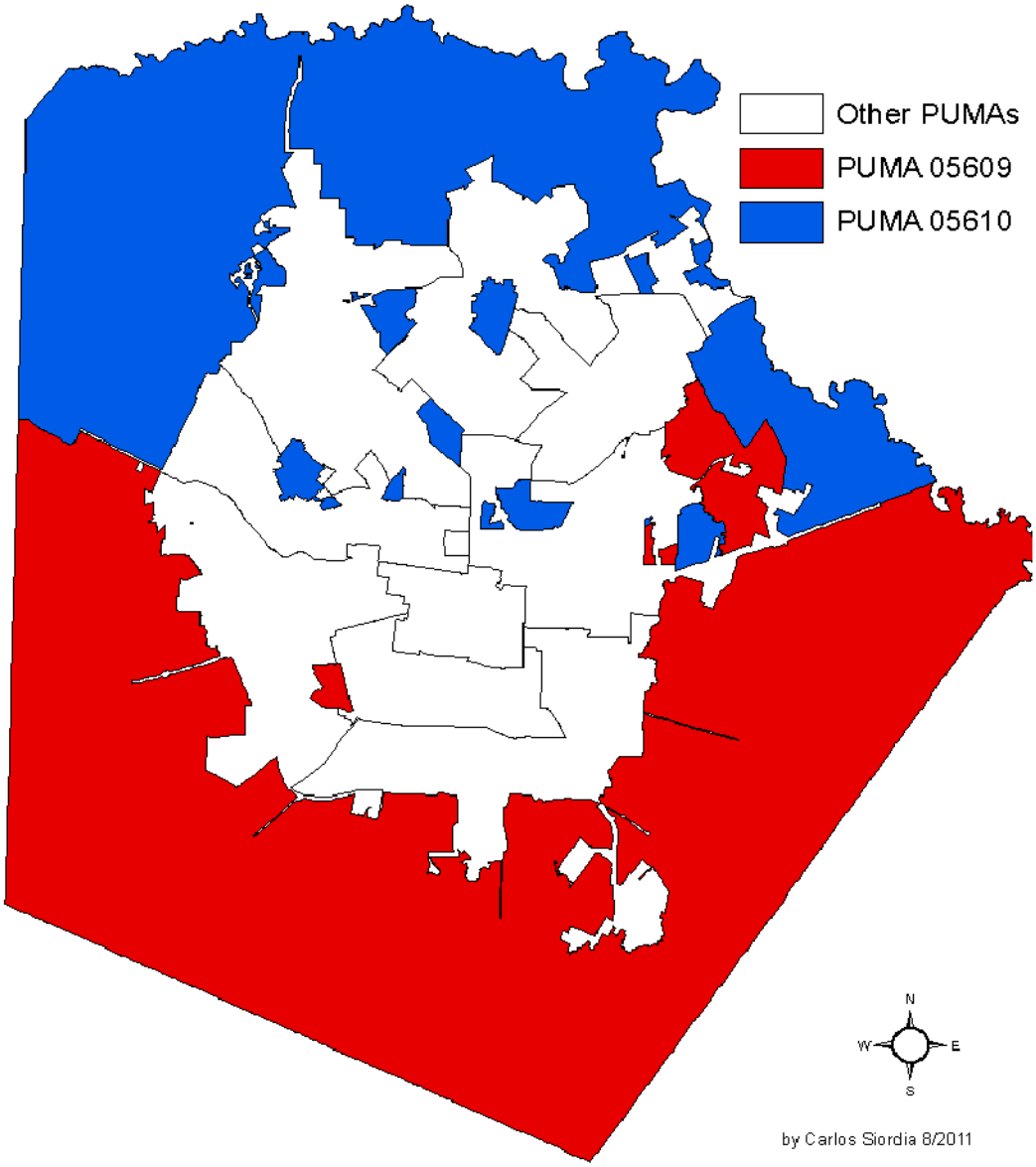


Figure 2. PUMA Fragmentation Results for Harris County, TX

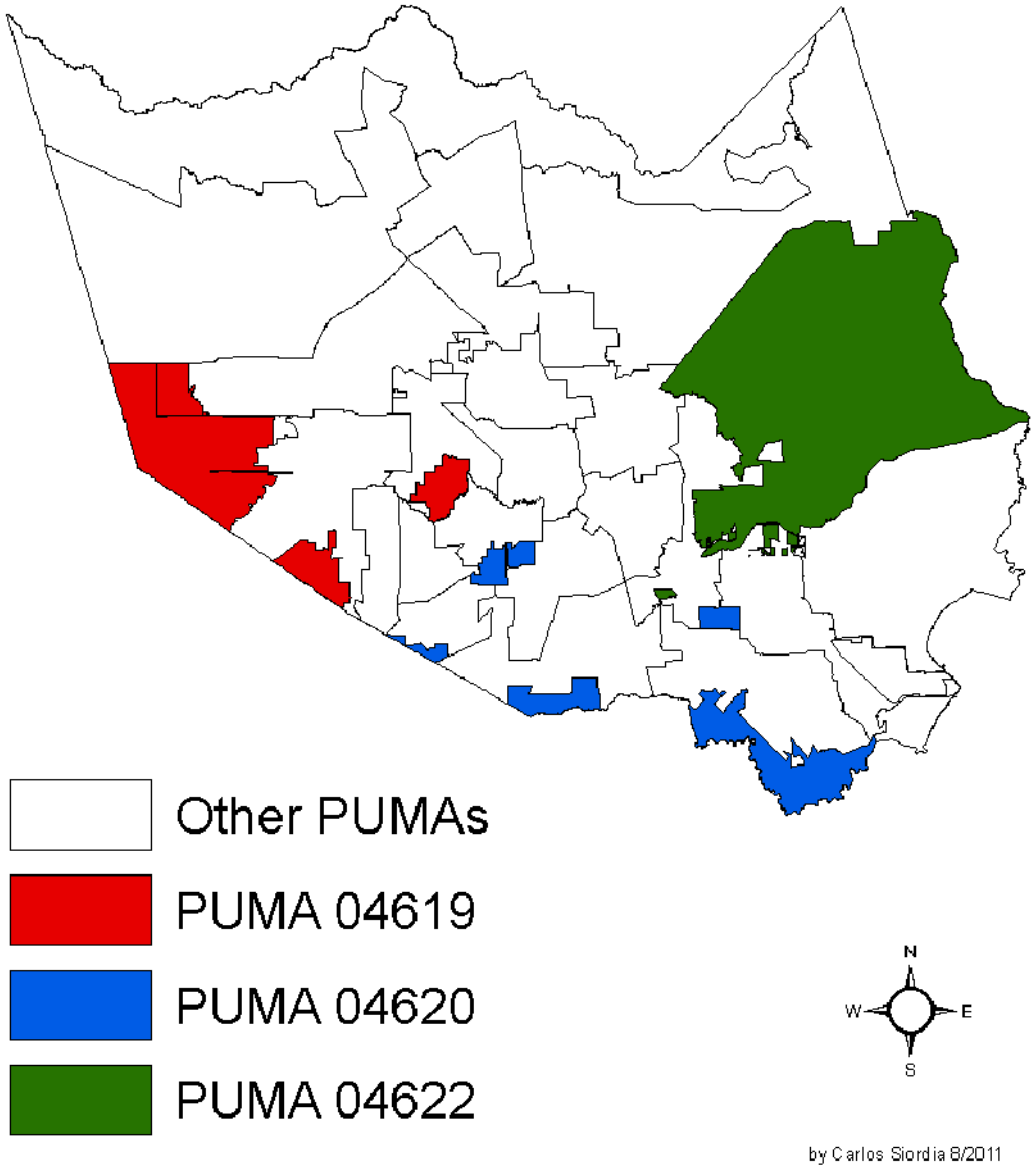
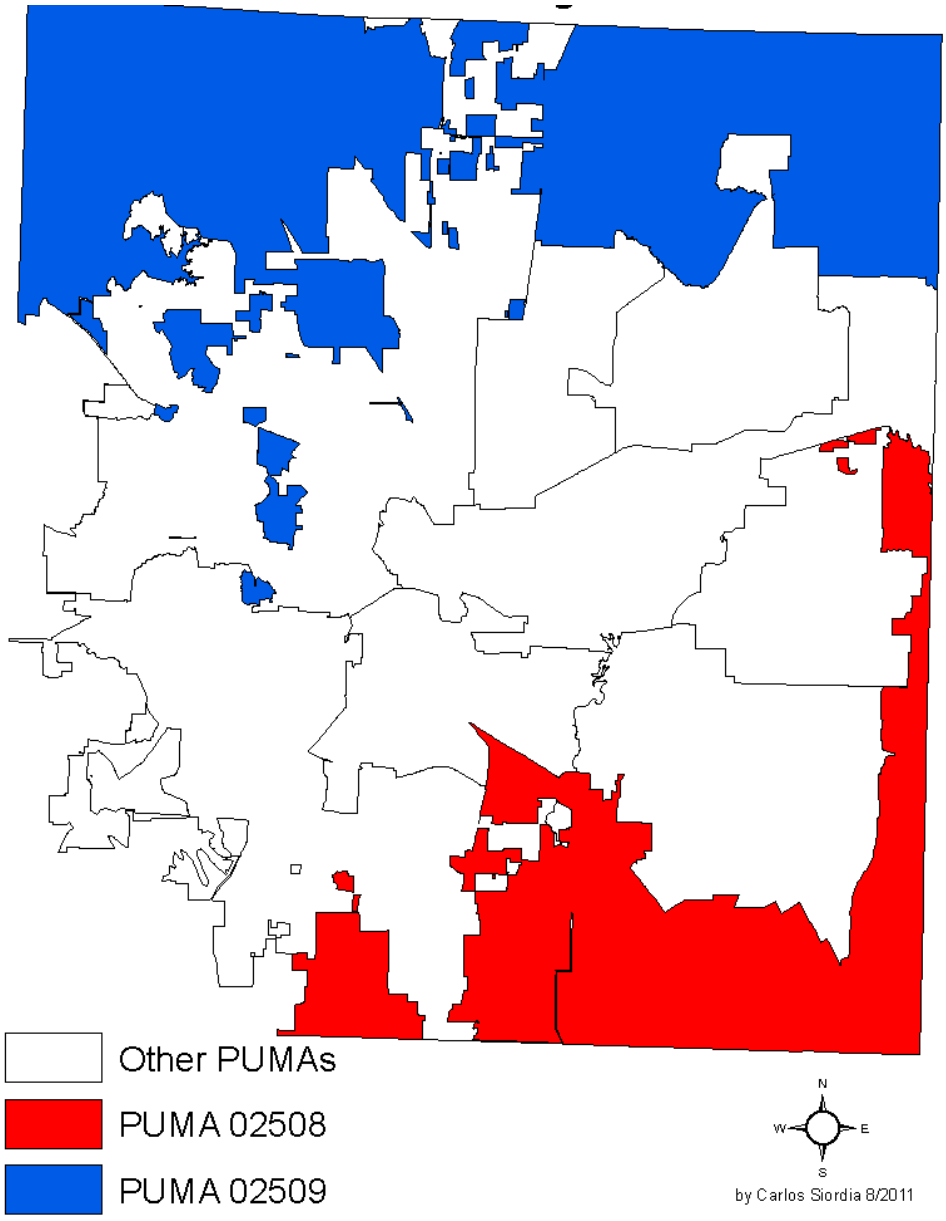


Figure 3. PUMA Fragmentation Results for Tarrant County, TX



a significant role in how PUMA polygons are delineated (for a more detail discussion on the quantity, degree, and location of PUMA fragmentation see Siordia and Wunneburger 2012).

CONCLUSIONS

We now give a summary of the problem and its importance, followed by one possible "administrative" solution. As stated in full length at the introduction of the paper, the problem is that PUMA polygon fragmentation is present. We have made the case that this is important because the problem challenges fundamental theoretical assumptions — which by extension weakens existing "socio-spatiological" approaches. Our figures have provided the evidence that PUMA

fragmentation is present. We have identified and discussed the three elements of fragmentation: quantity, distance, and size of each fragment. Each of these individually and in combination pose important research and policy questions that if answered, could help advance our understanding of spatial analysis and offer policy recommendations for improving PUMA delineation.

There are some limitations with our research. For example, we do not provide an analysis of how polygon centroids creating geospatial mismatches affect empirical findings. We discuss below why we omitted an investigation into this issue. It is important that future research develop techniques to investigate the influence multi-part polygons may have in biasing empirical results. Please note that multi-part polygons may in fact be present in other widely used Census polygons (e.g., tracts) —since fragmentations primarily occur because Census blocks are the element from which all other census geographies are built. As we have explained throughout, polygon discontiguity in spatial modeling may have several empirical and theoretical implications. For example, what do centroids derived from multi-part polygons and which are located outside all fragments mean in macro-level analyses of spatial nonstationarity? Techniques like spatial clustering and geographically weighted regression compute adjacency matrices that are said to capture how a human phenomenon varies as a function of geographical location. We believe geospatially mismatched centroids have the potential to influence statistical analyses and challenge the significance of findings.

What is a quantitative solution to handling polygon discontiguity? Existing research has shown how dissolving geospatially mismatched polygons on a common attribute can help (Grubesic and Matisziw 2006). In their research, Grubesic and Matisziw (2006) dissolve five-digit zip code tabulation area polygons split by water features on a common attribute identification to create contiguous polygons. Their main goal was to develop topologically rectified zip code tabulation areas to eliminate double entries in their data file—allowing for the creation of a more realistic adjacency matrix used in their spatial statistical analysis.

Table 1 Coefficient of Polygon Fragmentation by State

State	CPF_i	State	CPF_i
OK	3.24	MO	0.81
NV	2.78	NH	0.76
NC	2.44	MN	0.68
TX	2.23	NM	0.56
VT	2.08	PA	0.54
KS	1.98	NJ	0.41
WI	1.88	GA	0.40
ID	1.85	NY	0.35
ME	1.67	IN	0.35
CA	1.64	CT	0.33
AL	1.39	FL	0.33
TN	1.32	OR	0.31
CO	1.31	AZ	0.23
IA	1.31	VA	0.20
OH	1.28	MD	0.19
SC	1.23	MA	0.16
MI	1.22	MS	0.00
SD	1.19	UT	0.00
RI	1.19	NE	0.00
LA	0.93	WV	0.00
WA	0.91	MT	0.00
AR	0.88	DE	0.00
IL	0.86	DC	0.00
KY	0.83	ND	0.00
		WY	0.00

Our discussion has concentrated on giving evidence of polygon discontiguity in spatial research. We assume its presence affects empirical findings and have thusly made the argument by discussing what we believe is a fundamental assumption (polygon contiguity) in some statistical spatial techniques. We have not formally

investigated if and how the use of multi-part polygons affects empirical results for one main reason: in order to investigate the effect of discontiguous polygons on spatial modeling, we would have to first create a computational geometry solution and then choose (or create) a statistical technique for comparing the two results.

In the first stage, an algorithm would have to be created to dissolve multi-part polygons. This algorithm would ideally be capable of detecting discontiguous single features in vector files. After detecting multi-part polygons, the algorithm would then have to go through a series of theoretically informed decisions of how to dissolve the polygon. From Figure 1, the algorithm would have to have a set of rules on how to dissolve PUMA 4619: should all neighboring polygons be included?; should polygons enclosed by the fragments be the only ones included in the dissolve?; should more advanced techniques, like creating buffers, be used to identify polygons to be used in the dissolve? This hypothetical computational geometry algorithm would then be able to create a dataset which would only contain contiguous polygons and from which geospatially matched centroids could be derived. Even if such an algorithm were executed, a new challenge would be created: how would you justify that spatial attributes from dissolved polygons are as informative or equal in theoretical terms as measures derived from their smaller counterparts?; how could results from spatial models using two different sets of polygons be compared?; what technique could be employed to compare the influence of multi-part polygons after they have been dissolved using complex computational geometry algorithms?

Are there any non-quantitative solutions to eradicating multi-part polygons? We have one policy/administrative solution. In order for academic and policy researchers to better benefit from PUMS data, government stockholders (the Census, the SDCs, and all sub-SDCs partners) should be engaged to accomplish our suggested solution. The Census Bureau relies upon the expertise of data users such as demographers, economists, and regional experts to identify the areas or regions to include within specific PUMAs. In particular, the Census encourages SDCs to rely upon the expertise of data users to make these kinds of important decisions. We, the data users, hereby formally request all federal and local governmental agencies to undertake an evaluation of how PUMA fragmentation can be avoided. Ideally, we would recommend the formation of criteria that prohibits the production of fragmented PUMAs. The Census has already moved in the right direction with the 2010 guidelines. We propose PUMA boundaries be established through procedures that take into account the physical adjacency of the building blocks used in the formation of a recommended PUMA. We believe PUMA fragmentation can be mitigated by redefining procedures, set out by governmental agencies, to include a spatially aware agenda that visually works with block polygons to build non-fragmented PUMA geographies. In fulfilling this request, PUMAs can be made more useful to data users and will improve the quality and relevancy of PUMA-based research.

The theoretical, computational, and software advances in social geography must be complimented by advances in publicly available geographically referenced data. To our knowledge, this is the first paper to ever be published in an academic journal that clearly shows and discusses PUMA discontiguity and its implications for investigations using PUMS data. In particular, we feel our discussion of how the fundamental premise in the measurement of location necessitates both decipherable *and* contiguous geographical polygons is important and unique. We hope our exceptional discovery contributes to theory in spatial demography, as researchers in the field continue to grow in partnership with theoretical, statistical, data, and software advances.

Endnotes

1.For example see http://www.census.gov/did/www/sahie/methods/2000/estimates.html.

2.For example see http://www.census.gov/did/www/sahie/methods/2000/estimates.html

3.Further information on ACS PUMS data is provided

on the Census Bureau website: http://www.census.gov/acs/www/Downloads/handbooks/ACSPUMS.pdf.

4.Further details available at : http://www.census.gov/popest/data/historical/challenges.html.

5.Further details at Census Bureau website: http://www.census.gov/did/www/saipe/.

6.Further information about PUMS is provided on the Census Bureau website: http://www.census.gov/main/www/pums.html.

7.Information on PUMA delineation criterion can be found at: http://www.census.gov/geo/puma /2010_puma_guidelines.pdf.

8.The information in this paragraph was sourced and adapted from email correspondence with Vince Osier, the Brach Chief of the Geographic Standards & Criteria Brach in the Geography Division of the U.S. Census Bureau, Washington, DC.

9.Particulars on 2000 PUMA criteria can be found at: http://www.census.gov/geo/puma/puma_guide.pdf.

10.For a full history PUMS and PUMAs please visit: http://www.census.gov/geo/puma/puma_history.pdf.

11.Details on geographic terms and concepts are provided at: http://www.census.gov/geo/www/2010census/GTC_10.pdf.

12.Discussion available at: http://www.census.gov/geo/puma/FAQ_version2.pdf.

References

Alba, Richard. John Logan, Amy Lutz, and Brian Stults. 2002. "Only English by the Third Generation? Loss and Preservation of the Mother Tongue among the Grandchildren of contemporary Immigrants." *Demography*, 36(3):467-484.

Arnott, Richard. 2008. Economic Theory and the Spatial Mismatch Hypothesis. *Urban Studies*, 45; 2179-2202.

Barrios, Thomas, Rebecca Diamond, Guido W. Imbens, and Michal Kolesar. 2010. "Clustering, Spatial Correlations and Randomization Inference." *National Bureau of Economic Research*, NBER Working Paper No. 15760, Issued in February 2010.

Cova TJ, Church RL. 2000. Contiguity constraints for single-region site search problems. *Geographical Analysis*, 32(4):306-329.

de Berg, Mark, Otfried Cheong, Marc van Dreveld, and Mark Overmars. 2008. *Computational Geometry: Algorithms and Applications*. Springer, Verlag Berlin Heidelberg.

ESRI. 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.
ESRI. 1998. *Shapefile Technical Description: An ESRI White Paper*. California: Redlands.

Fotheringham, A. Steward, Chris Brunsdon, and Martin E. Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. West Sussex, UK: John Wiley.

Grubesic, Tony H., and Timothy C. Matisziw. 2006. On the Use of ZIP Codes and ZIP Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data. *International Journal of Health Geographics*, 5(58):1-15.

Jacquez GM. (2004). Current practices in the spatial analysis of cancer: flies in the ointment. *International Journal of Health Geographics,* 3(22):2-22.

Kain, John F. 1968. Housing Segregation, Negro Employment, and Metropolitan Decentralization. *Quaterly Journal of Economics*, 82(2): 175-197.

Mitchell, Andy. 2005. *The ESRI Guide to GIS Analysis, Volume 1: Geographic Patterns and Relationships and Zeroing In: Geographic Information Systems at Work in the Community*. ESRI Press, US.

Raphael, Steven, and Michael A. Stoll. 2010. Job Sprawl and the Suburbanization of Poverty. Metropolital Policy Program at Brookings, Metropolitan Opportunity Series, March: 1-21.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods. Second Edition*. Thousand Oaks, California: Sage Publications.

Reamer, Andrew D. 2010. "Surveying for Dollars: The Role of the American Community Survey in the Geographic Distribution of Federal Funds." Metropolital Policy Program at Brookings, Brookings, July: 1-18.

Saporito, Salvatore, Jana M. Chavers, Laura C. Nixon, Megan R. McQuiddy. 2007. "From here to there: Methods of allocating data between census geography and socially meaningful areas." *Social Science Research*, 36:897–920.

Siordia, Carlos, and Farias, R.A. 2013. A Multilevel Analysis on Latino's Economic Inequality: A Test of the Minority Group Threat Theory. Pp. 65-79 in the Economic Status volume of the Hispanic Population, edited by Richard Verdugo, in-press.

Rivers, Kerri L. 2006. *Using Data from the American Community Survey to Strengthen State Policies*. The Working Poor Families Project, Policy Brief, Winter.

Tobler, W 1970, "A Computer Movie Simulating Urban Growth in the Detroit Region", *Economic Geography*, vol.46, p.234-240.

U. S. Census Bureau. 2007. *2007 TIGER/Line Shapefiles*. Technical Documentation prepared by the U.S. Census Bureau, Washington, DC.

Wong, David W.S. 1998. "Spatial Patterns of Ethnic Integration in the United States." *Professional Geographer*, 50(1):13-30.

Yu, Zhou and Dowell Myers. 2007. "Convergence or Divergence in Los Angeles: Three Distinctive Patterns of Immigrant Residential Assimilation." *Social Science Research*, 36:254-285.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.